International Journal of Research and Applications
Jan-Mar © 2014 Transactions; 1(1): 31-35

ISSN (online): 2349-0020
http://ijraonline.com

COMPUTER SCIENCE ORIGINAL ARTICLE

ADVANCE TEXT STEGANOGRAPHY ALGORITHMS: AN OVERVIEW

K. Aditya Kumar [1], Dr. Suresh Pabboju [2] and Neela Megha Shyam Desai [3]

[1] Research Scholar, Osmania University, Hyderabad, A.P. India.

[2] Head, Department of IT , Chaitanya Bharathi Institute of Technology, Hyderabad, A.P. India

[3] Head, Department of CSE, Krishna Murthy Institute of Technology and Engineering, Hyd. , A.P. India

[1] kommera.aditya@gmail.com, [2] plpsuresh@gmail.com, [3] neelamsdesai@gmail.com

ABSTRACT

Today, in the digital age, any type of data, such as text, images, and audio, can be digitized, stored indefinitely, and transmitted at high speeds. Notwithstanding these advantages, digital data also have a downside. They are easy to access illegally, tamper with, and copy. The issue of information security has become increasingly important with the development of computer and expanding its use in different areas of life and work. One of the grounds discussed in information security is the exchange of information through the cover media. In this, different methods such as cryptography, steganography, etc., have been used. The main goal of steganography is to hide information in the other cover media so that other person will not notice the presence of the information. In steganography, the existence of the information in the sources will not be noticed at all. This paper gives a brief idea about various algorithms of text steganography.

1. INTRODUCTION

Steganography is derived from a finding by Johannes Trithemus (1462-1516) entitled "Steganographia", meaning "covered writing". Steganography is the art and science of hiding a message within a message without drawing any suspicion to others so that the message can only be detected by its intended recipient. Cryptography and Steganography are ways of secure data transfer over the Internet. Cryptography scrambles a message to conceal its contents; steganography conceals the existence of a message.

Steganography can be classified into image, text, audio and video steganography depending on the cover media used to embed secret data (Fig 1).

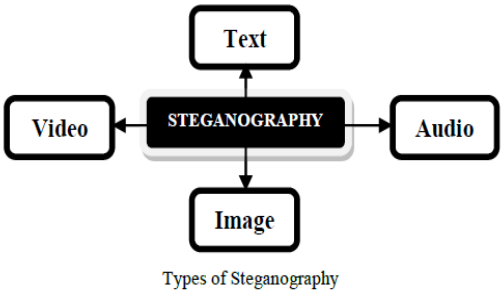

Fig 1: Types of Steganography

The advantage of using steganography is to conceal information. The transmission of messages is transparent to any given viewer. Messages can be concealed in different formats that are undetectable and unreadable to the human eye. Steganographic technologies are very important in Internet privacy today. With the use of steganography and encryption, corporations, governments, and law enforcement agencies can communicate secretly.

2. BASIC FEATURES

The steganographic methods have various strengths and weaknesses and this section discusses the few features of data hiding algorithms.

2.1 EMBEDDING CAPACITY:

Data are hidden (or embedded) in a larger volume of data called a cover or a carrier. The cover can be text, image, audio, or video. The amount of data that can be hidden in a cover, compared to the size of the cover is termed as Embedding capacity (also known as *payload*). This feature can be measured numerically in units of bit-per-bit (bpb). A steganographic algorithm with small embedding capacity may have other good features such as robustness, so it may be

the ideal choice when only a small amount of data, such as a short message, has to be hidden.

2.2 INVISIBILITY:
Any data hidden in a cover causes it to be modified. The measure of the amount of distortion to the cover is termed as Invisibility (also termed perceptual transparency or algorithm quality). A large embedding capacity is useless if it causes large distortions to the cover. Invisibility is a qualitative feature and it cannot be measured numerically. The best way of measuring it is to present several observers with the cover before and after the embedding. If no one can tell the difference between the covers, the steganographic algorithm is judged highly invisible. Invisibility is therefore tied to human visual or auditory perception.

2.3 UNDETECTABILITY:
An attacker may be able to detect the presence of hidden data in a given file by computing certain statistical properties of the file and comparing them to what is expected in that type of file. For example, if a particular image is examined and is found to have a significantly different pixel distribution; it may raise suspicion and lead to further scrutiny. Thus, a good steganographic method should not change the statistical properties of the cover file. This property is termed Undetectability and is different from invisibility because it does not depend on human perception.

2.4 ROBUSTNESS:
Even after the cover has been subjected to various changes as a result of lossy compression and decompression or of certain types of processing (such as conversion to analog and back to digital), the measure of the ability of an algorithm to retain the data embedded in the cover is termed as robustness. Most steganographic algorithms embed data in an image, and images may be subject to image processing operations (such as filtering, color changes, rotating, cropping, resampling, and sharpening). Robustness is especially important when the hidden data consist of copyright or ownership information (the so-called watermark). A user may compress such an image with a lossy compression method, and then decompress it in an attempt to destroy any hidden watermarks.

3. TEXT STEGANOGRAPHY
Text steganography involves anything like changing the format of an existing text, changing words within a text, generating random character sequences. Due to deficiency of redundant information which is present in image, audio or a video file, text steganography is believed to be the trickiest technique. In text documents, we can hide information by introducing changes in the structure of the document without making a notable change in the concerned output. Unperceivable changes can be made to an image or an audio file, but, in text files, even an additional letter or punctuation can be marked by a casual reader. Storing text file require less memory and its faster as well as easier communication makes it preferable to other types of steganographic methods. Text steganography can be broadly classified into three types (Fig 2): Format based Random and Statistical generation, Linguistic methods.

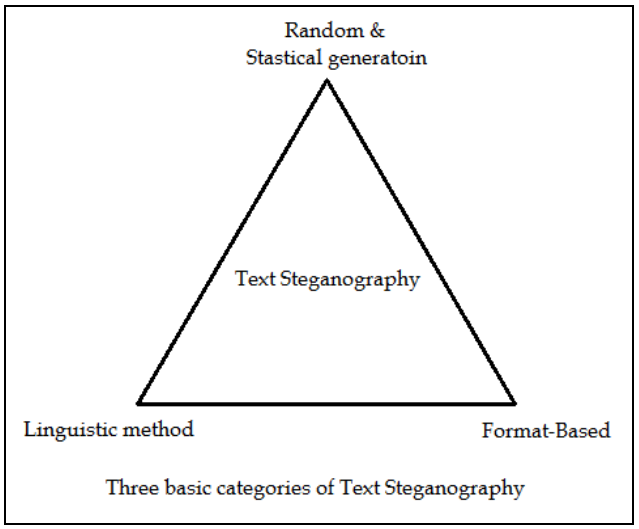

Fig 2: Categories of Text Steganography

3.1 FORMAT BASED METHODS
Format based methods involve altering physically the format of text to conceal the information. The disadvantage of this method is, if the stego file is opened with a word processor, misspellings and extra white spaces will get detected. Changed fonts sizes can arouse suspicion to a human reader. And, if the original plaintext is available, comparing this plaintext with the suspected steganographic text would make manipulated parts of the text quite visible.

3.2 RANDOM AND STATISTICAL GENERATION

In order to overcome the problem of comparison with a known plaintext, steganographers often choose to generate their own cover texts. The methods used are - concealing information in random looking sequence of characters, the statistical properties of word length and letter frequencies are used in order to create words which will appear to have same statistical properties as actual words in the given language.

3.3 LINGUISTIC STEGANOGRAPHY

Linguistic steganography specifically considers the linguistic properties of generated and modified text, and in many cases, uses linguistic structure as the space in which messages are hidden. CFG create tree structure which can be used for concealing the bits where left branch represents '0' and right branch corresponds to '1'. A grammar in GNF can also be used where the first choice in a production represents bit 0 and the second choice represents bit 1. Some drawbacks of using this method are - a small grammar will lead to lot of text repetition, although the text is syntactically flawless, but there is a lack of semantic structure. The result is a string of sentences which have no relation to one another.

4. APPROACHES

Some of the popular approaches of text steganography are:

4.1 Line Shift

In this method, secret message is hidden by vertically shifting the text lines 1/300 of an inch up or down. The marked line helps in detecting the direction of movement of the marked line. To hide bit 0, a line is shifted up and to hide bit 1, the line is shifted down. Determination of whether the line has been shifted up or down is done by measuring the distance of the centroid of marked line and its control lines. The disadvantage of this approach is if the text is retyped (or) if a character recognition program (OCR) is used, the hidden information would get destroyed. Also, by using special instruments of distance assessment, the distances can be observed.

4.2 Word Shift

In this method, by shifting words horizontally and by changing distance between words, information is hidden in the text, i.e. left or right to represent bit 0 or 1 respectively. This method is acceptable for texts where the distance between words is varying. Words shift are detected using correlation methods. This method can be identified less, because change of distance between words to fill a line is quite common. Disadvantage of this approach is, if someone knows the algorithm of distances, using the difference in the distances one can obtain the hidden text by comparing the stego text with the algorithm. Also, retyping or using OCR programs destroys the hidden information.

4.3 White Steg

This technique uses white spaces for hiding a secret message [6]. There are three methods of hiding data using white spaces - Inter Sentence Spacing, End of Line Spaces, Inter Word Spacing technique.
In Inter Sentence Spacing, to hide a bit 0, we place single space and to hide bit 1 we place two spaces at the end of each terminating character. In End of Line Spaces, fixed number of spaces is inserted at the end of each line. For example, two spaces to encode one bit per line, four spaces to encode two bits and so on. In Inter Word Spacing technique, bit 0 is represented by one space after a word and bit 1 is represented by two spaces after a word. But, inconsistent use of white space is not transparent.

4.4 Spam Text

HTML and XML files can also be used to hide bits. Bit 0 is interpreted if there is different starting and closing tags and bit 1 is interpreted if single tag is used for starting and closing. In another technique, lack of space in a tag represents bit 0 and placing a space inside a tag represents bit 1.

4.5 Syntactic Method

This technique uses punctuation marks such as full stop (.), comma (,), etc. to hide bits 0 and 1. The problem with this method is that it requires identification of correct places to insert punctuation marks.

4.6 Word Mapping

This technique encrypts a secret message using genetic operator crossover and then embeds the resulting cipher text, taking two bits at a time, in a cover file by inserting blank spaces between words of even or odd length using a certain mapping technique. The embedding positions are saved in another file and transmitted to the receiver along with the stego object.

4.7 CSS (Cascading Style Sheet)

This technique encrypts a message using RSA public key cryptosystem and cipher text is then embedded in a Cascading Style Sheet (CSS) by using End of Line on each CSS style properties, exactly after a semicolon. Bit 0 is embedded by placing a space after a semicolon and bit 1 is embedded by placing a tab space after a semicolon.

4.8 Mixed-Case Font

In this method, the information can be hidden in English text using the letters as carriers. This approach will insert one character within each 7 letters [7] [8]. So the hiding capacity will be very high compared to other text steganography methods. This method was tested on some text to compute the capacity of the hiding and to check its advantages. It will attract no attention as it will be thought as a type of these new cool fonts (Fig 3).

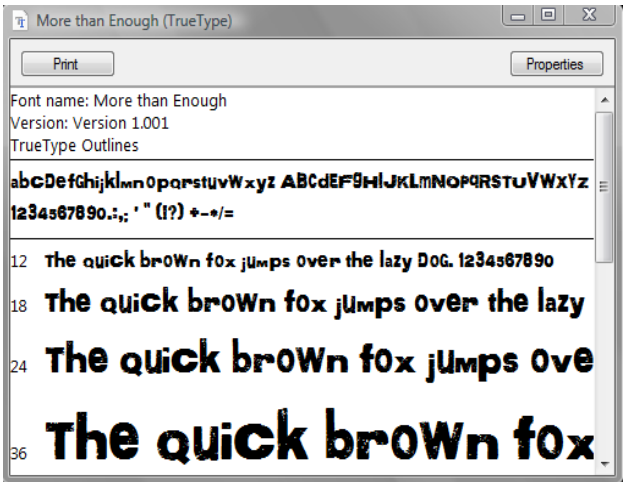

Fig 3: Mixed-Case Font

4.9 SMS-Texting

SMS-Texting language is a combination of abbreviated words used in SMS. By using full form of word or its abbreviated form, we can hide binary data. A codebook containing words and their corresponding abbreviated forms is made. To hide bit 0, full form of the word is used and to hide bit 1, abbreviated form of word is used.

4.10 Feature Coding

In feature coding, secret message is hidden by altering one or more features of the text. A parser examines a document and picks out all the features that it can use to hide the information. For example, point in letters i and j can be displaced, length of strike in letters f and t can be changed, or by extending or shortening height of letters b, d, h, etc. Disadvantage of this method is that the hidden content would get destroyed, if an OCR program is used or if retyping is done.

4.11 SSCE

(Secret Steganographic Code for Embedding)

This technique first encrypts a message using SSCE table and by using a certain mapping technique embeds the cipher text in a cover file by inserting articles a (or) an with the non specific nouns in English language[3]. The embedding positions are encrypted using the same SSCE table and saved in another file which is transmitted to the receiver securely along with the stego file.

4.12 Cricket Match Scorecard

In this method, data is hidden in a cricket match scorecard by pre-appending a meaningless zero before a number to represent bit 1 and leaving the number as it is to represent bit 0.

4.13 Missing letter puzzle

Missing letter puzzle is a puzzle comprising of a collection of words with one or more letters missing in each word. A letter, at some position in a word, is missed by replacing it with a question mark. The puzzle is solved by replacing each question mark by an appropriate letter in each word so as to make the words meaningful. Words in a puzzle can be of different length and different domains, i.e., they can be terminologies of various fields or can be proper nouns or a combination of both.

Each character of secret message is hidden in a word of certain length by missing one or two letters in it. Hint is also given with some words. Since the length of words depends on the decimal (ASCII) value of characters to be hidden, the words are dynamically generated and there is no pre-determined cover file.

4.14 Hiding Data in Wordlist

This method conceals a message in a list of words without using any special character. Each character is hidden in a word of certain length. The starting letter of word is determined by masking sum of the digits of ASCII value of the character to an English alphabet. If sum of the digits is 1, then starting letter of word will be ‘a’; if it is 2, then ‘b’ and so forth. Since the length and staring letter of words depends on the decimal value of the embedded characters, cover is dynamically generated.

4.15 Hiding Data in Paragraphs

This approach makes use of a pre-determined cover file which can be any meaningful piece of English text and can be drawn from any source (For example, a paragraph from a newspaper/book). The approach works by hiding a message using start and end letter of the words of a cover file. This approach works on the binary value of a character. After converting the cipher text to a stream of bits, each bit is hidden by picking a word from the cover file and using either the start or the end letter of that word depending on the bit to be concealed. Bit 0 or 1 is hidden by reading a word, sequentially, from the cover file and including the starting letter or the end letter, respectively, of the word in the stego key. A word having same start and end letter is skipped. Since no change is made to the cover, the cover file and its corresponding stego file are exactly the same.

5. CONCLUSIONS

This paper has provided an overview of steganography, which by definition literally means "covered writing". The basic methods of text steganography are Format-based, random and statistical generation, linguistic methods, are discussed. And some methods discussed for hiding data in text are – Line Shift, Word Shift, SSCE, White Steg, etc.

REFERENCES

1. http://www.springer.com/978-0-387-00311-5 Data privacy and security, Salomon D. 2003, XIV, 465 p. 122, hardcover ISBN: 978-0-387-00311-5.
2. F. A. P. Petitcolas, R.J. Anderson, and M. G. Kuhn, "Information hiding- a survey," In *Proceedings of IEEE*, vol.87, pp. 1062-1078, 1999.
3. Text Steganography using Article Mapping Technique (AMT) and SSCE Indradip Banerjee, Souvik Bhattacharyya and Gautam Sanyal.
4. Information Hiding: A New Approach in Text Steganography - L. Y. POR, B. Delina.
5. TextSteganographic Approaches: A Comparison - Monika Agarwal.
6. L.Y.Por, T. F. Ang, and B. Delina, "WhiteSteg- a new scheme in information hiding using text steganography," *WSEAS Transactions on Computers*, vol.7, no.6, pp. 735-745, 2008.
7. StegChat: A Synonym-Substitution Based Algorithm for Text Steganography - Joseph Gardiner Supervisor: Dr. Shishir Nagaraja.
8. New text steganographic technique by using mixed case font – Abdelmgeid Alim Ali, Al – Hussien Seddik Saad.
9. Steganography an Art of Hiding Data - Shashikala Channalli, Ajay Jadhav.